Proceedings of the 2007 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2007)

Cost-Sensitive Fuzzy Classification for Medical Diagnosis

G. Schaefer
School of Engineering and Applied Science
Aston University
U.K.

T. Nakashima, Y. Yokota and H. Ishibuchi
Department of Computer Science
Osaka Prefecture University
Japan

***Abstract*— Medical diagnosis essentially represents a pattern classification problem: based on a certain input an expert arrives at a diagnosis which often takes on a binary form, i.e. the patient suffering from a certain disease or not. A lot of research has focussed on computer assisted diagnosis where objective measurements are passed to a classifier algorithm which then proposes diagnostic output based on a previous learning process. However, these classifiers put equal emphasis on a learning patterns irrespective of the class they belong to. In this paper we apply a fuzzy rule-based classification system to medical diagnosis. Importantly, we extend the classifier to incorporate a concept of cost which can be used to emphasize those cases that signify illness as it is usually more costly to incorrectly diagnose such a patient as being healthy. Experimental results on various medical datasets confirm the usefulness and efficacy of our approach.**

Keywords: medical diagnosis, pattern classification, cost-sensitive classification, fuzzy classification

I. Introduction

Medical diagnosis is often regarded as a pattern classification problem: based on a certain input the task is to assign it to one of a set of classes, where the number of classes is often two, e.g. malignant vs. benign, disease vs. no disease. The task of a pattern classification system is to assign as many input samples as possible to the correct class whereas the behaviour of the classifier is often optimised through the learning of some ground truth data. Conventional classifiers treat each sample of this learning set equally, yet in medical diagnosis this is often not desirable as different classes are associated with different costs. While the misdiagnosis of a malignant case as being benign can be very costly (e.g. when the time for effective treatment has passed) the mistaking a benign case as malignant (though of course it should be avoided) will involve relatively lower costs such as some further tests.

In this paper we present a cost-sensitive approach to medical diagnosis based on fuzzy rule-based classification. While fuzzy rule-based systems have been mainly employed for control problems [1] more recently they have also been applied to pattern classification problems [2], [3]. We modify a fuzzy rule-based classifier to incorporate the concept of weight which can be considered as the cost of an input pattern being misclassified. The pattern classification problem is thus reformulated as a cost minimisation problem. Based on experimental results on three standard medical datasets (Wisconsin breast cancer, heart disease, and diabetes data sets from the UCI machine learning repository) we demonstrate the efficacy of our approach. We also show that the application of a learning algorithm can further improve the classification performance of our classifier.

The rest of the paper is organised as follows: Section II covers in detail fuzzy rule-based classification systems. Our cost-sensitive fuzzy classifier is explained in Section III while Section IV provides a learning algorithm that can be applied to boost the classification rate. Experimental results on various datasets are given in Section V. Section VI concludes the paper.

II. Fuzzy rule based classification

Let us assume that our pattern classification problem is an n-dimensional problem with M classes (in medical diagnosis M is typically 2) and m given training patterns $\mathbf{x}_p = (x_{p1}, x_{p2}, \ldots, x_{pn})$, $p = 1, 2, \ldots, m$. Without loss of generality, we assume each attribute of the given training patterns to be normalised into the unit interval $[0, 1]$; that is, the pattern space is an n-dimensional unit hypercube $[0, 1]^n$. In this study we use fuzzy if-then rules of the following type as a base of our fuzzy rule-based classification systems:

$$\begin{array}{l}\text{Rule } R_j\text{: If } x_1 \text{ is } A_{j1} \text{ and } \ldots \text{ and } x_n \text{ is } A_{jn} \\ \text{then Class } C_j \text{ with } CF_j, \quad j = 1, 2, \ldots, N,\end{array} \tag{1}$$

where R_j is the label of the j-th fuzzy if-then rule, $A_{j1}, \ldots, A_{jn}$ are antecedent fuzzy sets on the unit interval $[0, 1]$, C_j is the consequent class (i.e. one of the M given classes), and CF_j is the grade of certainty of the fuzzy if-then rule R_j. As antecedent fuzzy sets we use triangular fuzzy sets as in Figure 1 where we show a partition of the unit interval into a number of fuzzy sets.

Our fuzzy classification system consists of N fuzzy if-then rules each of which has a form as in Equation (1). There are two steps in the generation of the rules: specification of antecedent part, and determination of consequent class C_j and grade of certainty CF_j. The antecedent part of a rule is specified manually. Then the consequent part (i.e. consequent class and the grade of certainty) is determined from the given training patterns [4]. In [5] it is shown that the use of the grade of certainty in fuzzy if-then rules allows us to generate comprehensible fuzzy rule-based classification systems with high classification performance.

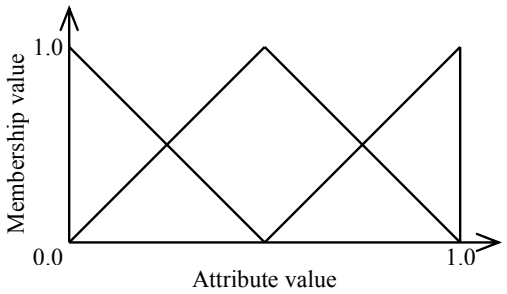

Fig. 1. Membership function.

A. Fuzzy rule generation

Let us assume that m training patterns $\mathbf{x}_p = (x_{p1}, \ldots, x_{pn})$, $p = 1, \ldots, m$, are given for an n-dimensional C-class pattern classification problem. The consequent class C_j and the grade of certainty CF_j of the if-then rule are determined in the following manner:

Step 1: Calculate $\beta_{\text{Class } h}(j)$ for Class h as

$$\beta_{\text{Class } h}(j) = \sum_{\mathbf{x}_p \in \text{Class } h} \mu_j(\mathbf{x}_p), \tag{2}$$

where

$$\mu_j(\mathbf{x}_p) = \mu_{j1}(x_{p1}) \cdot \ldots \cdot \mu_{jn}(x_{pn}), \tag{3}$$

and $\mu_{jn}(\cdot)$ is the membership function of the fuzzy set A_{jn}. In this paper, we use triangular fuzzy sets as in Figure 1.

Step 2: Find Class $\hat{h}$ that has the maximum value of $\beta_{\text{Class } h}(j)$:

$$\beta_{\text{Class } \hat{h}}(j) = \max_{1 \le k \le C} \{\beta_{\text{Class } k}(j)\}. \tag{4}$$

If two or more classes take the maximum value, the consequent class C_j of the rule R_j cannot be determined uniquely. In this case, we specify C_j as $C_j = \phi$. If a single class $\hat{h}$ takes the maximum value, let C_j be Class $\hat{h}$.

The grade of certainty CF_j is determined as

$$CF_j = \frac{\beta_{\text{Class } \hat{h}}(j) - \bar{\beta}}{\sum_h \beta_{\text{Class } h}(j)} \tag{5}$$

with

$$\bar{\beta} = \frac{\sum_{h \neq \hat{h}} \beta_{\text{Class } h}(j)}{c - 1}. \tag{6}$$

B. Fuzzy reasoning

Using the rule generation procedure outlined above we can generate N fuzzy if-then rules as in Equation (1). After both the consequent class C_j and the grade of certainty CF_j are determined for all N rules, a new pattern $\mathbf{x} = (x_1, \ldots, x_n)$ can be classified by the following procedure:

Step 1: Calculate $\alpha_{\text{Class } h}(\mathbf{x})$ for Class h, $j = 1, \ldots, C$, as

$$\alpha_{\text{Class } h}(\mathbf{x}) = \max\{\mu_j(\mathbf{x}) \cdot CF_j | C_j = h\}, \tag{7}$$

Step 2: Find Class h' that has the maximum value of $\alpha_{\text{Class } h}(\mathbf{x})$:

$$\alpha_{\text{Class } h'}(\mathbf{x}) = \max_{1 \le k \le C} \{\alpha_{\text{Class } k}(\mathbf{x})\}. \tag{8}$$

If two or more classes take the maximum value, then the classification of $\mathbf{x}$ is rejected (i.e. $\mathbf{x}$ is left as an unclassifiable pattern), otherwise we assign $\mathbf{x}$ to Class h'.

III. Cost-sensitive fuzzy classification

The standard fuzzy rule-based classifier as detailed above treats each class and hence each sample equally. In medical diagnosis however this is often not desirable. Misdiagnosing a malignant case as benign should be penalised more than diagnosing healthy patients as having a certain disease. While in the first case the result might be that of late treatment in the best and missing of the treatable time in the worst scenario, the latter case will usually involve some further tests which should then identify the misdiagnosis. Clearly, the costs involved in the first case will exceed those of the latter.

We therefore wish to develop a classifier that incorporates this and reformulate the pattern classification problem as a cost minimisation problem. The concept of a weight is introduced for each training pattern in order to handle this situation where the weight of an input pattern can be viewed as the cost of misclassifying it. Fuzzy if-then rules are generated by considering the cost as well as the compatibility of training patterns.

In order to incorporate the concept of weight/cost, we modify Equation 2 of the fuzzy rule generation to

$$\beta_{\text{Class } h}(j) = \sum_{\mathbf{x}_p \in \text{Class } h} \mu_j(\mathbf{x}_p) \cdot \omega_p \tag{9}$$

where ω_p is the cost associated with training pattern p. A suitable overall cost function can be defined as

$$\texttt{Cost}(S) = \sum_{p=1}^{m} \omega_p \cdot z_p(S), \tag{10}$$

where m is the number of training patterns, ω_p is the weight/cost of the training pattern $\mathbf{x}_p$, and $z_p(S)$ is a binary variable set according to the classification result of the training pattern $\mathbf{x}_p$ by S: $z_p(S) = 0$ if $\mathbf{x}_p$ is correctly classified by S, and $z_p(S) = 1$ otherwise (i.e. $\mathbf{x}_p$ is misclassified or rejected). We use this cost function as a performance measure as well as classification rate.

IV. Learning algorithm

A learning method that adjusts the grades of certainty CF_j can be employed to achieve improved classification performance [6]. It is based on an error-correction learning approach where the adjustment occurs when classification of training patterns is not successful. When a training pattern is correctly classified we do not adjust the grade of certainty. The main idea of the learning method is to adjust the degree of certainty CF_j of two fuzzy if-then rules: We decrease the degree of certainty of a fuzzy if-then rule that misclassifies

Dataset	# of attributes	total # of cases	# malignant	# benign
breast cancer [7]	9	683	239	444
heart disease [8]	13	270	120	150
diabetes [9]	8	768	268	500

TABLE I

STATISTICS OF DATASETS.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	18	97.49	98.65	233	6	6	438
	cost-based	15	98.33	98.42	235	4	7	437
	cost+learning $\eta = 0.2, K = 50$	2	100.00	99.55	239	0	2	442
	cost+learning $\eta = 0.5, K = 20$	2	100.00	99.55	239	0	2	443
1:5	conventional	36	97.49	98.65	233	6	6	438
	cost-based	21	99.16	97.52	237	2	11	433
	cost+learning $\eta = 0.2, K = 50$	2	100.00	99.55	239	0	2	442
	cost+learning $\eta = 0.5, K = 20$	2	100.00	99.55	239	0	2	443

TABLE II

10-CV RESULTS ON BREAST CANCER TRAINING DATA.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	39	94.14	97.52	225	14	11	433
	cost-based	37	95.40	96.66	228	11	15	429
	cost+learning $\eta = 0.2, K = 50$	37	95.40	96.66	228	11	15	429
	cost+learning $\eta = 0.5, K = 20$	37	95.40	96.66	228	11	15	429
1:5	conventional	81	94.14	97.52	225	14	11	433
	cost-based	56	96.65	96.40	231	8	16	428
	cost+learning $\eta = 0.2, K = 50$	56	96.65	96.40	231	8	16	428
	cost+learning $\eta = 0.5, K = 20$	56	96.65	96.40	231	8	16	428

TABLE III

10-CV RESULTS ON BREAST CANCER TEST DATA.

a training pattern and in turn increase that of a fuzzy if-then rule that is supposed to correctly classify the training pattern.

Let us assume that we have generated fuzzy if-then rules by the rule-generation procedure detailed in Section II-A. We also assume that a fuzzy if-then rule R_j misclassifies a training pattern $\mathbf{x}_p$. That is, R_j is used to classify $\mathbf{x}_p$ from Class c^* by using Equation (8) but the consequent class C_j does not agree with the true class of the training pattern $\mathbf{x}$. Let R_* be the fuzzy if-then rule that is selected by Equation (7). That is, R_* has the maximum value of $\alpha_{\text{Class } c^*}(\mathbf{x}_p)$ among those fuzzy if-then rules with Class c^* but does not have the maximum value among all generated fuzzy if-then rules. The proposed learning method adjusts the grades of certainty of R_j and R_* as follows:

$$CF_j^{new} = CF_j^{old} - \eta \cdot \omega_p \cdot CF_j^{old}, \tag{11}$$

$$CF_*^{new} = CF_*^{old} - \eta \cdot \omega_p \cdot (1 - CF_*^{old}), \tag{12}$$

where ω_p is the weight of the training pattern $\mathbf{x}_p$, and η is a positive constant value. We assume that $0 \leq \eta \leq 1$.

One epoch of the proposed learning method involves examining all given training patterns. Thus there will be $2m$ adjustments of fuzzy if-then rules if all m training patterns are misclassified. The learning process is summarised as follows:

Step 1: Generate fuzzy if-then rules from m given training patterns by the procedure in Section II-A.
Step 2: Set K as $K = 1$.
Step 3: Set p as $p = 1$.
Step 4: Classify $\mathbf{x}_p$ by using the generated fuzzy if-then rules in Step 1.
Step 5: If $\mathbf{x}_p$ is misclassified, adjust the grades of certainty using Equations (11) and (12). Otherwise no rules are adjusted.
Step 6: If $p < m$, let $p := p + 1$ and go to Step 4. Otherwise go to Step 7.
Step 7: If K reaches a pre-specified value, stop the learning procedure. Otherwise let $K := K+1$ and go to Step 3.

Note that K in the above learning procedure corresponds to the number of epochs.

V. EXPERIMENTAL RESULTS

In order to evaluate our proposed cost-sensitive fuzzy classifier we tested it on several standard medical classification datasets, in particular a breast cancer dataset [7], a dataset of heart patients [8], and the data of a study into diabetes [9]. Sizes and distribution of classes of all three datasets are given in Table I from where we can see that all three constitute two-class problems, with benign and malignant as the two target

classes.

For all three datasets we used a standard fuzzy rule-based classifier as described in Section II to obtain a baseline benchmark to compare our algorithms to. We then applied our proposed cost-based classifier with two different cost setting: a ratio of 1:2 between benign and malignant cases and a ratio of 1:5 (though both are probably still conservative estimates). We also used the learning algorithms detailed in Section IV to improve upon the classification performance of our cost-based classification system. We investigate two different sets of parameters for the learning algorithm: a slower learning method with $\eta = 0.2$ and $K = 50$ and a faster approach with $\eta = 0.5$ and $K = 20$. In all experiments we divide each attribute uniformly into three triangular fuzzy sets as shown in Figure 1. In order to arrive at statistically meaningful results, in all cases we perform 10-fold cross-validation where the patterns are split into ten disjoint subsets and each subset is in turn used as an unseen test set while the other nine sets are used for training the classifier. We report the results in terms of sensitivity defined as $SE = \frac{TP}{TP+FN}$ and specificity defined as $SP = \frac{TN}{TN+FP}$ where TP, TN, FP, and FN correspond to true positives, true negatives, false positives, and false negatives respectively. All results are given as the average 10-CV scores for both training and unseen test data.

Let's now inspect each dataset in more detail. The Wisconsin breast cancer dataset [7] is a collection of 9 cytological attributes such as clump thickness, uniformity of cell size and shape, etc. for 444 benign and 239 malignant cases. Classification results for the conventional fuzzy classifier, our cost-based variation, and the cost-based classifier after learning are shown in Tables II and III for training and test data respectively.

On the training data, the conventional classifier achieves a sensitivity and specificity of 97.49% and 98.65% respectively which corresponds to 6 false negatives and 6 false positives. Our cost-based approach improves upon this by correctly identifying two more malignant cases (while increasing the number of false positives by 1). A further dramatic improvement is observed after we apply the learning algorithm achieving a sensitivity of 100% with a specificity of 99.55%. On the test data the difference between the algorithms is still significant though not as pronounced as for the training data. Here, 3 more malignant cases are identified for a cost setting of 1:2 while 6 more cancer patients are detected with a cost setting of 1:5. The effect of the misclassification costs on the performance is hence readily observable here; also in all cases the cost-based variations produce lower overall costs compared to the conventional approach.

The heart diagnosis dataset contains 13 attributes which were derived from an initial set of 75 [8]. Apart from patient information such as age and sex, the attributes contain, among others, information on blood pressure, cholesterol, and blood sugar. Of the 270 patients 120 were diagnosed with a heart disease. Experimental results on all classifiers are given in Tables IV and V.

Again, significant improvement are being made through the application of the cost-based classifiers on the training data. Perfect classification results (i.e. 100% sensitivity and specificity) are achieved with the cost-based classifier after learning with a 1:2 cost setting. Performance on test data is significantly worse compared to the training data. This suggests that for this dataset is is fairly difficult to extract the correct rules and to generalise from a given training set. Also, the cost-based classifier performs fairly similar to the conventional algorithm on test data.

The third dataset we investigated is a diabetes database collected by the National Institute of Diabetes and Digestive Kidney Diseases [9]. Among the 8 attributes are indicators such as the patient's age, blood pressure, body mass index and others. Of the 768 patients 268 were tested positive for diabetes and the remaining 500 negative. We report the experimental results in Tables VI and VII.

Looking at these results we can first observe that the conventional classifier performs fairly poorly on this dataset, even on the training set. Even though a specificity of 96.40% is achieved the sensitivity here is 40.67%. We can also see that our cost-based classifier significantly improves upon that, achieving a sensitivity of up to 99.63% when combined with learning. On the other hand the specificity for these sets drops, which is not surprising given the aim of the classifier. In all cases the total costs for the complete dataset are well below those obtained from the conventional algorithm, in particular for the 1:5 cost ratio, both for training and test sets.

VI. Conclusions

In medical diagnostic classification systems, classification performance is not always the only indicator for assessing classifiers. Rather, misclassification costs should be taken into account as well, as usually misclassifying a malignant case will prove much more costly than misclassifying a benign one. In this paper we have applied a cost-sensitive fuzzy rule-based classifier, which emphasises the importance of those classes which have high misclassification costs, to various medical diagnostic classification datasets. We have also applied a learning algorithm to further boost the classification results. Experimental results have shown that our cost-based approaches perform better than conventional classifiers under the assumed conditions.

References

[1] C.C. Lee. 1990, "Fuzzy logic in control systems: Fuzzy logic controller part I and part II," *IEEE Trans. Systems, Man and Cybernetics*, vol. 20, pp. 404–435, 1990.

[2] K. Nozaki, H. Ishibuchi, and H. Tanaka, "Adaptive fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Systems*, vol. 4, no. 3, pp. 238–250, 1996.

[3] H. Ishibuchi and T. Nakashima, "Performance evaluation of fuzzy classifier systems for multi-dimensional pattern classification problems," *IEEE Trans. Systems, Man and Cybernetics - Part B: Cybernetics*, vol. 29, pp. 601–618, 1999.

[4] H. Ishibuchi, K. Nozaki, and H. Tanaka, "Distributed representation of fuzzy rules and its application to pattern classification," *Fuzzy Sets and Systems*, vol. 52, no. 1, pp. 21–32, 1992.

[5] H. Ishibuchi and T. Nakashima, "Effect of rule weights in fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Systems*, vol. 9, no. 4, pp. 506–515, 2001.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	10	96.67	98.67	116	4	2	148
	cost-based	7	99.17	96.67	119	1	7	145
	cost+learning $\eta = 0.2, K = 50$	0	100.00	100.00	120	0	0	150
	cost+learning $\eta = 0.5, K = 20$	0	100.00	100.00	120	0	0	150
1:5	conventional	22	96.67	98.67	116	4	2	148
	cost-based	12	100.00	92.00	120	0	12	138
	cost+learning $\eta = 0.2, K = 50$	3	100.00	98.00	120	0	3	147
	cost+learning $\eta = 0.5, K = 20$	3	100.00	98.00	120	0	3	147

TABLE IV

10-CV RESULTS ON HEART TRAINING DATA.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	152	55.83	69.33	67	53	46	104
	cost-based	154	56.67	66.67	68	52	50	100
	cost+learning $\eta = 0.2, K = 50$	155	56.67	66.00	68	52	51	99
	cost+learning $\eta = 0.5, K = 20$	155	56.67	66.00	68	52	51	99
1:5	conventional	311	55.83	69.33	67	53	46	104
	cost-based	306	58.33	62.67	70	50	56	94
	cost+learning $\eta = 0.2, K = 50$	305	58.33	63.33	70	50	51	99
	cost+learning $\eta = 0.5, K = 20$	305	58.33	63.33	70	50	55	95

TABLE V

10-CV RESULTS ON HEART TEST DATA.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	336	40.67	96.40	109	159	18	482
	cost-based	275	68.66	78.60	184	84	107	393
	cost+learning $\eta = 0.2, K = 50$	238	82.09	71.60	220	48	142	358
	cost+learning $\eta = 0.5, K = 20$	301	98.88	41.00	265	3	295	205
1:5	conventional	813	40.67	96.40	109	159	18	482
	cost-based	362	98.51	31.60	264	4	342	158
	cost+learning $\eta = 0.2, K = 50$	296	99.63	41.80	267	1	291	209
	cost+learning $\eta = 0.5, K = 20$	298	99.25	42.40	266	2	288	212

TABLE VI

10-CV RESULTS ON DIABETES TRAINING DATA.

	Classifier	tot. cost	SE [%]	SP [%]	TP	FN	FP	TN
1:2	conventional	372	35.45	94.80	95	173	26	474
	cost-based	300	64.92	77.60	174	94	112	388
	cost+learning $\eta = 0.2, K = 50$	313	70.52	69.00	189	79	155	345
	cost+learning $\eta = 0.5, K = 20$	366	88.88	38.88	238	30	306	194
1:5	conventional	891	35.45	94.80	95	173	26	104
	cost-based	399	95.52	32.20	256	12	339	161
	cost+learning $\eta = 0.2, K = 50$	457	89.55	36.60	240	28	317	183
	cost+learning $\eta = 0.5, K = 20$	443	90.30	37.40	242	26	313	187

TABLE VII

10-CV RESULTS ON DIABETES TEST DATA.

[6] T. Nakashima, Y. Yokota, H. Ishibuchi, and G. Schaefer, "Learning fuzzy if-then rules for pattern classification with weighted training patterns," in *4th Conference of the European Society for Fuzzy Logic and Technology*, 2005, pp. 1064–1069.

[7] W.H. Wolberg and O.L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proceedings of the National Academy of Sciences*, vol. 87, pp. 9193–9196, 1990.

[8] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, and V. Froelicher, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *American journal of Cardiology*, vol. 64, pp. 304–310, 1989.

[9] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes, "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in *Proc. Symp. Computer Applications and Medial Cares*, 1983, pp. 422–425.